Modeling Subject-Specific Formant Transition Patterns in /ɑʃɑ/ Sequences

Martine Toda[+, ++], Shinji Maeda[++], Michaël Aron[+++] and Marie-Odile Berger[+++]

[+]Université Paris III / CNRS UMR 7018
[++]Telecom-ParisTech / CNRS-LTCI UMR 5141
[+++] LORIA/INRIA Nancy-Grant Est
E-mail: martinetoda@yahoo.co.jp

Abstract

Careful observation reveals that the formant transition patterns in a particular VCV sequence vary according to subjects. It is the case in /aʃa/ uttered by French speakers. This variation can be explained by the realization of acoustically equivalent but differently articulated /ʃ/ or /a/ segments and their combination.

This paper reports an acoustic simulation experiment aiming at replicating this phenomenon. The results are compared to simultaneous ultrasound tongue contour data and discussed in the perspective of speech inversion.

1 Introduction

In a previous study, we have identified two subject-specific varieties of French /ʃ/ [1]. After the sustained utterance of /ʃ/ imaged by magnetic resonance (MRI), four subjects out of seven used the tongue 'position' strategy, consisting of backing the tongue without significantly changing its shape with respect to the contrasting sibilant /s/. The remaining three subjects used the tongue 'shape' strategy, resulting generally in a palatalized articulation of /ʃ/ due to a domed tongue dorsum. Only the 'position' variety tended to be accompanied by a lip protrusion (with respect to /s/ taken as reference), and its sublingual cavity was often deeper due to a protruded or raised tongue tip.

Those two types of /ʃ/ are however likely to produce similar frication noise peaks [2], thus being potentially a source of uncertainty in articulatory inversion. Moreover, contrary to what we would have expected, the F2 and F3 transition patterns in the /aʃa/ token did not vary consistently in function of the /ʃ/-strategy. The subject-specific articulation of the surrounding vowel /a/ has been suggested to contribute to this complex articulatory-to-acoustic relationship.

Figure 1. *'Position'-strategy /ʃ/ (left) vs. 'shape'-strategy /ʃ/ (right). Vocal-tract contours of /ʃ/ (black) and /s/ (gray; adapted from [1]).*

The present paper reports a vowel-fricative-vowel (VFV) simulation experiment where both /a/ and /ʃ/ models are manipulated in order to fit the /a/ and /ʃ/ targets in a subject's recorded utterance. The formant transition patterns are then discussed with respect to the their cavity affiliation during targets.

2 Method

2.1 VFV synthesis

The target /aʃa/ utterance (Figure 2) is modeled by means of time-domain transmission-line analog acoustic simulation program, vcvSynt [3]. The target vocal-tract (VT) area function is given as input for each segment and its shape is interpolated section-by-

section (section length = 5 mm) during the VF and FV transitions.

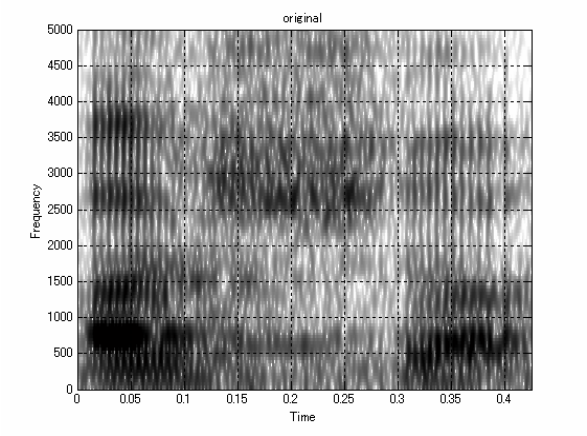

Figure 2. *Target /aʃa/ utterance produced by subject FH.*

The fundamental frequency, glottal parameters as well as timing were set to suitable values in order to match the target utterance. The glottal source switches to a frication noise source when the tightest point of constriction in the mouth region becomes narrower than the glottal opening. Then the vocal-tract is excited by a noise source (pink noise) which location is specified by the user (actually set at the incisors).

2.2 Optimization of /a/ target

It is known that the vocal tract shape vary according to the subjects both in length as well as in vertical (pharyngeal) vs. horizontal (oral) ratio [4]. According to the authors of this study, the inter-subject variation observed in the articulatory movements involved in vowel production shall be explained by these anatomical differences. Therefore, we defined the /a/ model (Figure 3) on articulatory continua so that the total length (from 13 to 20 cm) and the back/front cavity length ratio (from 0.3 to 1.7) would approximate the vocal tract of any subject. The configurations that gave the closest F1-F4 frequencies to those measured in the steady-state portion of the post-consonantal /a/ of the target utterance (in least squares; the same weight being allocated for F1 through F4 expressed in Hertz) were selected as the optimal /a/ configurations.

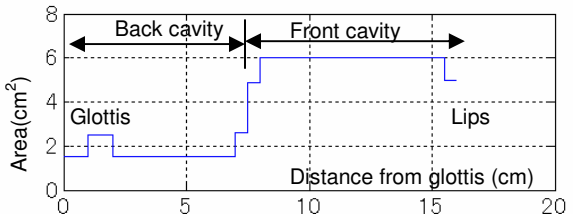

Figure 3. */a/ model specified by two parameters: total length and back/front cavity ratio.*

Actually two local minima were found: total length = 16 cm; back/front vocal-tract ratio = 0.9 and length = 16 cm; ratio = 1.4. The 'posterior' variety (ratio = 0.9) is probably more realistic with regard to the relative intensity of F3 and F4, in that the F3 bandwidth of the natural utterance tends to be wider than that of F4 (suggesting lip radiation, thus a front cavity affiliation).

2.3 Selection of appropriate /ʃ/ configurations

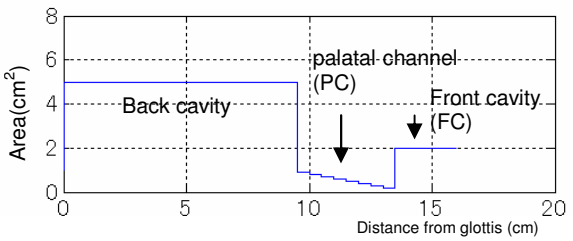

Figure 4. */ʃ/ model with variable palatal channel and front cavity lengths.*

The /ʃ/ model is made up of a back cavity, a constriction, and a front cavity (Figure 4). Its total length was set identical to the /a/ model (16 cm). The appropriate constriction (palatal channel) and front cavity lengths were chosen so that the lowest prominent peak of their transfer function (see [2]) matched that of the target utterance (2664 Hz). Among the possible candidates, some configurations ('apical': PC = 0.5 cm, FC = 4.5 cm; 'palatalized': PC = 5.5 cm, FC = 2 cm; thin plain lines in Figure 5) fitted better the noise spectrum of the target utterance (bold line), in that they possess secondary peaks around 5.5 and 6.5 kHz, either of which is assumed to be the three-quarter wave-length resonance of the front cavity viewed as a closed tube (apical configuration), or the 1 wave-length

resonance of the palatal channel considered as an open tube (palatalized configuration).

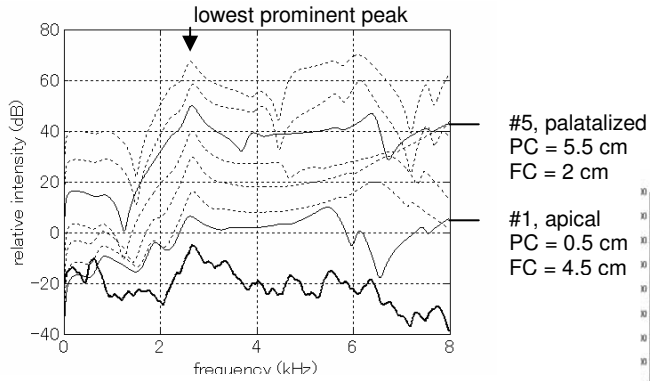

Figure 5. *Time-averaged spectrum of /ʃ / extracted from the target /aʃ a/ utterance (bottom thick line) and transfer functions (shifted by 20 dB one another) of /ʃ / configurations giving rise to a lowest prominent peak close to that of the natural utterance. The palatal channel lengthens and the front cavity shortens from bottom to top.*

3 Synthesis results

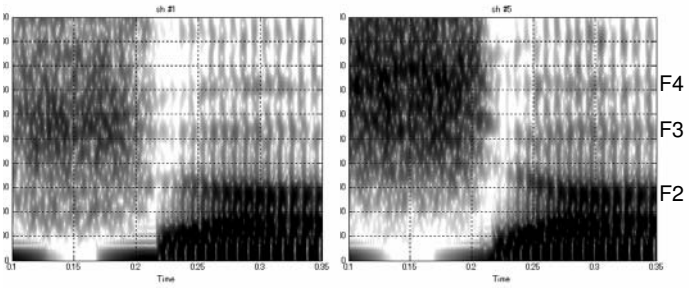

Figure 6. *Spectrograms of synthesized /aʃa/ sequences ([ʃa] transition shown). /a/ length = 16 cm, ratio = 0.9. Left: 'apical' /ʃ/; right: 'palatalized' /ʃ/.*

The VFV simulation result with a short constriction (apical) /ʃ/ model (Figure 6, left) agrees better with the target [ʃa] formant transition pattern. The F2 onset at the fricative-vowel junction is lower (pointing towards 1.6 kHz into the fricative instead of above 2 kHz in the palatalized /ʃ/), and the F3 is flat (instead of a slightly rising pattern in the palatalized /ʃ/). The 'anterior' variant of /a/ (ratio = 1.4) gives similar results (Figure 7 left, to be compared to Figure 6 right; by virtue of the acoustic law stating that formants never cross each other, similar transitions result when the formant pattern of the targets are the same, even though their cavity affiliation differ).

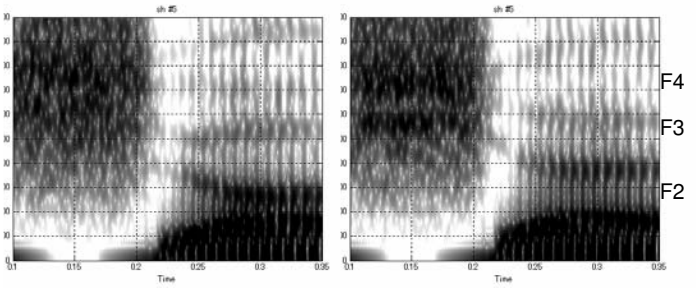

Figure 7. *More spectrograms of synthesized [ʃa] transition with 'palatalized' /ʃ/. Left: 'anterior' /a/ of length = 16 cm, ratio = 1.4; right: female-like 'posterior' /a/ of length = 13 cm, ratio = 0.5.*

When the same fricatives are combined to a female-like /a/ (total length of the fricative adjusted by shortening the back cavity), striking differences appear (Figure 7). With a palatalized /ʃ/ (#5), a male subject would exhibit a rising F3 and falling F4 towards the following vowel, whereas a female subject whose formants are more sparsely distributed will show a slightly falling F3 and a rising F4.

4 Interpreting formant transitions

The preceding sections described a VFV simulation experiment based on acoustic targets where vocal tract models served as interface. This procedure can be considered to be an articulatory inversion in ideal conditions (all the presumably useful acoustic cues are exploited, and all the technical problems are ignored). In this section, we aim at evaluating the likelihood of such a 'manual' inversion, and interpreting the acoustic mechanisms underlying the formant transitions.

In a male-like vocal tract (Figure 6), the F2 onset at the [ʃ-a] junction is not connected to a noise peak, thus indicating a resonance of the back cavity (which is not significantly excited by the noise). Everything

being equal, a palatalized articulation has a shorter back cavity (Figure 8, bottom). Then it makes sense that its resonances are higher in frequency compared to the apical variant.

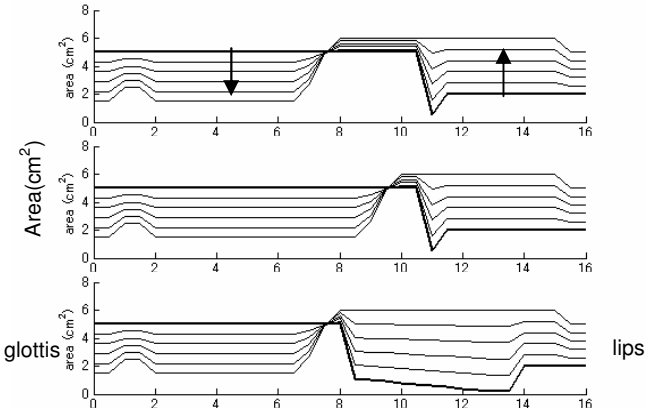

Figure 8. *Vocal tract area function during the transition from /ʃ/ (bold line) to /a/. Top: apical /ʃ/ (#1), /a/ with ratio = 0.9; middle: /ʃ/ #1, /a/ ratio = 1.5; bottom: palatalized /ʃ/ (#5), /a/ ratio = 0.9.*

In a female-like vocal tract, the back cavity of the palatalized variety is even shorter than the palatal channel itself, thus explaining the connection of F2 to the lowest prominent peak affiliated to the palatal channel, which produces the second lowest resonance after the Helmholtz resonance connected to F1 (Figure 7, right).

To summarize, an apical variety of /ʃ/ gave rise to a better F2 transition. The MRI tongue contour of sustained /ʃ/ (Figure 9) corroborate our prediction, by showing a tendency for this speaker to make a short constriction. In addition, during the [ʃ]-to-[a] transition, a portion of the vocal tract remains steady (Figure 8; the upper two panels are 'apical' /ʃ/). These steady segments do look compatible with the actual tongue contours (velar region is steady) extracted from ultrasound frames acquired simultaneously to the acoustic recordings (Figure 9; see [5] for details).

In conclusion, the appropriate /a/ and /ʃ/ models derived from the acoustics, in particular formant targets and transitions, permit to re-synthesize subject-specific instances of /aʃa/. A further step will consist in examining which transitional patterns are actually observed, and how they are conditioned by articulatory economy, or some other principle, that would be specific to their actual vocal tract.

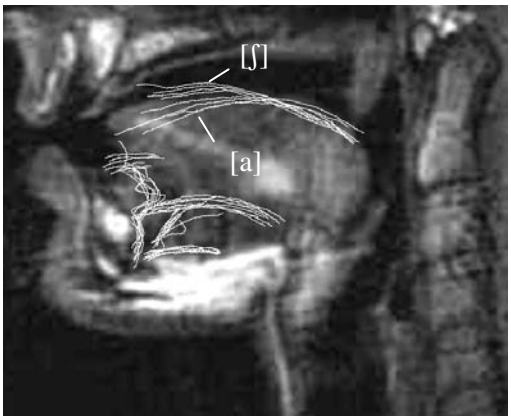

Figure 9. *Ultrasound tongue contours during the target /aʃa/ utterance superimposed onto the midsagittal MR image of sustained /ʃ/ (alignment cued by tongue textures); subject FH.*

6 Acknowledgement

The authors acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Sixth Framework Programme for Research of the European Commission, under FET-Open contract no. 021324 (**ASPI** project).

7 References

[1] M. Toda, 'Deux stratégies articulatoires pour la réalisation du contraste acoustique des sibilantes /s/ et /ʃ/ en français', *proc. XXIV^th Journées d'Étude de la Parole* (Dinard, France), 2006.

[2] M. Toda and S. Maeda, 'Quantal aspects of non anterior sibilant fricatives: a simulation study', *proc. 7^th ISSP* (Ubatuba, Bresil), 2006.

[3] S. Maeda, 'Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer', *proc. Sound patterns of connected speech description, models and explanation symposium* (Kiel, Germany), 1996.

[4] K. Honda, S. Maeda, M. Hashi, J. S. Dembowski, J. R. Westbury, 'Human palate and related structures: their articulatory consequences' *proc. 4^th ICSLP*, 1996.

[5] M. Aron, M.-O. Berger, E. Kerrien, 'Multimodal fusion of electromagnetic, ultrasound and MRI data for building an articulatory model', *proc. 8^th ISSP* (Strasbourg, France), 2008.